Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues

Atefeh Farzindar[*] [**], Mathieu Roche[***] [****]

[*] *NLP Technologies Inc., Montréal (Québec), Canada*
www.nlptechnologies.ca
farzindar@nlptechnologies.ca
[**] *Université de Montréal, Montréal (Québec), Canada*
[***] *UMR TETIS (Cirad, Irstea, AgroParisTech), Montpellier, France*
mathieu.roche@cirad.fr
[****] *LIRMM (CNRS, Université de Montpellier), Montpellier, France*
mathieu.roche@lirmm.fr

RÉSUMÉ. *Les réseaux sociaux intègrent un volume et une variété sans précédent de données textuelles. Leur analyse permet de mieux comprendre des comportements sociaux et certaines évolutions sociétales. L'étude des messages échangés, qui sont par nature complexes, représente de nouvelles problématiques pour le traitement automatique des langues (TAL). Dans ce contexte, cet article introductif au numéro spécial de la revue TAL présente les défis liés à l'infobésité des données issues des réseaux sociaux puis discute de l'utilisation des méthodes de TAL pour traiter le contenu textuel de ces nouveaux modes de communication.*

ABSTRACT. *Social networks incorporate an unprecedented amount and variety of textual data. The analysis of this information furthers our understanding of social behaviors and some trends. The study of inherently complex messages sent between users represents new problems for Natural Language Processing (NLP). In this context, the first article in this special issue of the TAL journal introduces the challenges of information overload from social networks, and discusses the use of NLP methods for processing the textual content of these new modes of communication.*

MOTS-CLÉS : *TAL, réseaux sociaux, analyse sémantique.*

KEYWORDS: *NLP, social networks, semantic analysis.*

1. Introduction

Les réseaux sociaux, structures dynamiques formées d'individus ou d'organisations, ont toujours joué un rôle majeur dans nos sociétés. Ils se sont développés et diversifiés avec le Web 2.0 qui ouvre la possibilité aux utilisateurs de créer et de partager du contenu par l'intermédiaire de multiples plates-formes (blogues, microblogues, wikis, sites de partage, etc.). Ces modes de communication sont de puissants outils collectifs où s'invente et s'expérimente le langage. De nouveaux sens sont alors associés à certains mots ou syntagmes et la création de mots ou de nouvelles structures syntaxiques se généralise. La création, la dissémination et le traitement du matériau textuel issu des réseaux sociaux sont discutés dans ce numéro spécial. Plus globalement, cet article et ce numéro spécial permettent de mettre en exergue une nouvelle manière de communiquer illustrée par (1) l'article « Code-Mixing in Social Media Text : The Last Language Identification Frontier ? » de Amitava Das et Björn Gambäck sélectionné parmi sept articles soumis et (2) l'article invité « Détection d'évènements à partir de Twitter » de Houssem Eddine Dridi et Guy Lapalme.

Pour traiter les masses de données issues des réseaux sociaux aujourd'hui disponibles (c'est-à-dire l'infobésité), la problématique de recherche du « Big Data » est classiquement mise en avant avec les trois V qui la caractérisent : volume, variété et vélocité. Cet article discute, dans un premier temps, de ces trois caractéristiques appliquées aux réseaux sociaux (section 2). Puis, dans le cadre de l'infobésité décrite de manière générale, nous étudierons, en section 3, la manière d'analyser le contenu des messages issus des réseaux sociaux par des méthodes de traitement automatique des langues (TAL). En effet, certaines métadonnées (par exemple, les hashtags) et les descripteurs linguistiques (ou unités lexicales) issus des messages constituent un socle solide pour l'analyse des réseaux sociaux. Ils permettent de mettre en avant différentes communautés socio-économiques, politiques, géographiques, etc. Par ailleurs, les descripteurs linguistiques sous forme de mots ou syntagmes permettent d'analyser avec précision les sentiments et opinions contenus dans les messages. Par exemple, les spécificités lexicales, graphiques voire syntaxiques (émoticônes, abréviations, répétition de caractères, etc.) véhiculent des informations précieuses pour l'analyse de sentiment (détection fine des émotions, identification de l'ironie, etc.).

2. Infobésité et analyse des réseaux sociaux

Au cœur de la structure des réseaux sociaux se trouvent des acteurs (personnes ou organisations) reliés entre eux par un ensemble de relations binaires (par exemple, liens ou interactions). Dans ce contexte, le but est de modéliser la structure d'un groupe social, en vue de déterminer l'influence qu'elle exerce sur d'autres variables, et d'assurer le suivi de son évolution. L'analyse sémantique des médias sociaux (ASMS) se définit comme l'art de comprendre comment on recourt aux réseaux sociaux pour générer du renseignement sociétal, stratégique, opérationnel ou tactique. Récemment, des ateliers tels que « L'analyse sémantique des médias sociaux » et « L'analyse linguistique dans les médias sociaux » de EACL 2012 (Farzindar et Inkpen, 2012),

NACL-HLT 2013 (Farzindar *et al.*, 2013) et EACL 2014 (Farzindar *et al.*, 2014) témoignent de l'intérêt grandissant à l'égard de l'impact des médias sociaux sur la vie quotidienne des individus, tant sur le plan personnel que professionnel. L'ASMS favorise la création d'outils et d'algorithmes visant à surveiller, à saisir et à analyser les données des médias sociaux qui sont volumineuses (section 2.1), produites en temps réel (section 2.2) et de nature hétérogène (section 2.3).

2.1. *Volume*

Un rapport publié par eMarketer (New Media Trend Watch, 2013) estimait qu'une personne sur quatre à l'échelle mondiale était susceptible d'utiliser les médias sociaux en 2013. Les statistiques sur les médias sociaux pour l'année 2012 révèlent que Facebook a dépassé la barre des huit cents millions d'utilisateurs actifs, dont deux cents millions de nouveaux adhérents au cours d'une seule année. La plate-forme Twitter, quant à elle, compte maintenant cent millions d'utilisateurs et LinkedIn, plus de soixante-quatre millions, en Amérique du Nord seulement (Digital Buzz, 2012). À titre d'exemple, plus de trois cent millions de tweets seraient envoyés à Twitter chaque jour (Tang *et al.*, 2014).

L'analyse et la veille de ce riche contenu sans cesse renouvelé donnent accès à une information précieuse que les médias traditionnels ne peuvent fournir (Melville *et al.*, 2009). L'analyse sémantique des médias sociaux a ouvert la voie à l'analyse de données volumineuses, discipline émergente inspirée de l'apprentissage automatique, de l'exploration de données, de la recherche documentaire, de la traduction automatique, du résumé automatique et du TAL plus globalement.

2.2. *Vélocité*

Les données issues des réseaux sociaux sont en général produites en temps réel. Par ailleurs les messages traitant d'un sujet commun peuvent véhiculer des émotions, des néologismes ou des rumeurs. Ces messages peuvent provenir de localisations différentes qu'il est nécessaire de prendre en compte dans le cadre de la vélocité des données.

Les médias sociaux soulèvent l'important problème de la recherche d'événements en temps réel et de la nécessité de les détecter (Farzindar et Wael, 2015). L'objectif de la recherche documentaire dynamique et de la recherche d'événements en temps réel est de mettre en place des stratégies de recherche efficaces à partir de différentes fonctionnalités qui tiennent compte de multiples dimensions, y compris les liens spatiaux et temporels (Gaio *et al.*, 2012 ; Moncla *et al.*, 2014). En outre, les discussions propres à un événements peuvent mêler, sur une période très courte, différents sujets parfois écrits en différentes langues. Ce point illustre la problématique liée à l'hétérogénéité des données qui est détaillée dans la section suivante.

2.3. *Variété*

L'importante quantité d'informations accessible dans les médias sociaux représente une manne de renseignements. Mais les textes, rédigés par des auteurs différents dans une variété de langues et de styles, n'adoptent, en général, aucune structure précise et se présentent sous une multitude de formats : blogues, microblogues, forums de discussion, clavardages, jeux en ligne, annotations, classements, commentaires et FAQ générées par des utilisateurs, etc. Les variations sur le plan du contenu et du style rendent l'analyse globale difficile. De manière concrète, les applications décrites ci-dessous montrent la variété des domaines et des tâches menées à partir des réseaux sociaux.

2.3.1. *Secteur industriel*

L'intérêt pour la surveillance de données extraites des médias sociaux est considérable dans le secteur industriel. En effet, ces données sont susceptibles d'aider en optimisant de manière importante l'efficacité de la veille stratégique. L'intégration de telles données aux systèmes de veille stratégique déjà en place permet aux entreprises d'atteindre différents objectifs, notamment concernant la stratégie de marque et la notoriété, la gestion des clients actuels et potentiels et l'amélioration du service à la clientèle. Le marketing en ligne, la recommandation de produits et la gestion de la réputation ne sont que quelques exemples d'applications concrètes de l'ASMS.

2.3.2. *Défense et sécurité nationale*

Ce secteur s'intéresse en particulier à l'étude de ce type de sources d'information pour comprendre différentes situations, procéder à l'analyse des sentiments d'un groupe de personnes partageant des intérêts communs et rester vigilant aux menaces potentielles dans les domaines cibles. Certaines méthodes d'extraction d'information (par exemple l'extraction des entités nommées et des liens entre ces dernières) à partir du Web 2.0 sont souvent développées pour analyser le contenu des réseaux sociaux au sein desquels évoluent des utilisateurs mais aussi des organisations. De telles informations offrent de précieux renseignements en matière de sécurité nationale.

2.3.3. *Soins de santé*

Les forums de discussion qui sont des espaces d'échanges asynchrones de messages textuels sont très prisés par certains patients. En effet, ils sont associés à un véritable espace de liberté du discours. Ainsi, l'utilisation de Twitter ou des forums comme des plates-formes de discussion sur des sujets tels que les maladies, les traitements, les médicaments ou les recommandations à l'intention des professionnels et des bénéficiaires (patients, familles et aidants) illustre bien la pertinence des médias sociaux dans ce domaine. Par ailleurs, dans ce contexte éminemment subjectif, la caractérisation et la compréhension des perceptions que les patients ont de leur maladie et du suivi médical représentent un enjeu sociétal particulièrement intéressant pour les professionnels de santé (Bringay *et al.*, 2014 ; Abdaoui *et al.*, 2014).

2.3.4. *Politique*

La veille des médias sociaux permet d'assurer le suivi des mentions faites par différents citoyens d'un pays ainsi que de l'opinion à l'égard d'un parti politique. Le nombre d'abonnés que compte un parti est essentiel au déroulement de sa campagne électorale. L'extraction d'opinions et le suivi des déclarations publiées sur les réseaux sociaux permettent à un parti politique de mieux saisir la teneur de certains événements, lui donnant ainsi l'occasion de s'ajuster pour améliorer ses positionnements politiques voire ses propositions (Bouillot *et al.*, 2012 ; Bakliwal *et al.*, 2013).

Face au volume, à la vélocité et à la variété des données textuelles issues des réseaux sociaux, les méthodes de TAL à appliquer et à proposer se révèlent cruciales. Les nouveaux défis adressés au TAL dans un tel contexte sont détaillés dans la section suivante.

3. Les défis liés au traitement du contenu des réseaux sociaux

L'information diffusée dans les médias sociaux, notamment dans les forums de discussion, les blogues et les gazouillis, est riche et dynamique. L'application des méthodes habituelles de TAL dans ce contexte ne se fait pas sans difficulté en raison du bruit et de l'orthographe « inhabituelle ». L'importance des médias sociaux émane du fait que chaque utilisateur est désormais un auteur potentiel et que le langage se rapproche davantage de sa réalité que d'une quelconque norme linguistique (Zhou et Hovy, 2006). Les blogues, les gazouillis et les mises à jour de statuts sont rédigés de manière informelle, sur le ton de la conversation, et ressemblent plus à un « état d'âme » qu'au travail réfléchi et révisé avec le soin habituellement attendu d'un média papier. Ce caractère informel engendre différents défis au domaine du TAL.

Les outils du TAL conçus pour les données traditionnelles se heurtent, par exemple, à l'emploi irrégulier, voire l'omission, de la ponctuation et des majuscules. Une telle situation complique la détection des limites d'une phrase qui constitue une tâche de base essentielle pour l'analyse des textes. Par ailleurs, l'utilisation de binettes, l'orthographe incorrecte ou inhabituelle et la multiplication d'abréviations populaires compliquent les tâches telles que la segmentation et l'étiquetage morphosyntaxique. Une adaptation des outils traditionnels est nécessaire pour prendre en compte les nouvelles variations comme la répétition des lettres (par exemple, *suuuuuper*) (Hangya *et al.*, 2013). Un autre obstacle à toute forme d'analyse syntaxique est la grammaticalité, ou plutôt son absence fréquente dans les médias sociaux (Kong *et al.*, 2014). En effet, les phrases fragmentées sont devenues la norme à l'instar des phrases complètes et le choix entre différents homophones semble arbitraire (par exemple, *c'est*, *ces*, *ses*).

Outre ces aspects liés aux spécificités lexicales voire syntaxiques du contenu des messages échangés sur les réseaux sociaux, ces derniers génèrent beaucoup plus de bruit que les médias dits traditionnels. En effet, les réseaux sociaux comportent un

nombre considérable de pourriels, de publicités et une importante quantité de contenus non sollicités, non pertinents ou dérangeants. En outre, une grande partie du contenu qualifié d'authentique et de légitime ne répond pas mieux aux besoins d'information, et est donc jugée non pertinente, comme l'illustrait bien l'étude rapportée dans (André *et al.*, 2012), visant à mesurer la valeur que les utilisateurs accordent à différents gazouillis. Des quarante mille évaluations de gazouillis recueillies, 36 % recevaient la mention « vaut la peine d'être lu » et 25 %, « ne vaut pas la peine d'être lu ». Les gazouillis qui attestent seulement de la présence d'un utilisateur sur la plate-forme (par exemple, *Alloooo Twitter !*) se sont vus attribuer la plus faible valeur. Cela souligne l'importance du prétraitement, visant à filtrer les pourriels et autres contenus non pertinents, et de la création de modèles de gestion du bruit efficaces, en vue du traitement du langage dans les médias sociaux.

De nombreux domaines d'application qui prennent en compte ces caractéristiques propres aux réseaux sociaux sont alors étudiés comme par exemple, le résumé automatique, la détection d'événements et l'analyse de sentiments. Ces trois domaines de recherche appliqués aux réseaux sociaux et caractérisés par les trois V du « Big Data » sont détaillés ci-dessous.

– Comme illustré en section 2.1, avec la présence de textes courts, bruités et en nombre important, les médias sociaux se prêtent difficilement aux approches de TAL comme le résumé automatique. À titre d'exemple, les gazouillis, avec leur limite de cent quarante caractères, sont plus pauvres sur le plan contextuel que les documents traditionnels. Aussi, la redondance est problématique dans une suite de gazouillis, en partie en raison de la fonction de partage. Les expériences présentées dans (Sharifi *et al.*, 2010) avec les techniques d'exploration de données visant à générer des résumés automatiques de sujets à la mode sur Twitter les ont amenés à identifier l'important problème posé par la redondance de l'information. En outre, l'information diffusée dans les médias sociaux est hautement dynamique et caractérisée par l'interaction entre différents participants. Si elle complexifie d'autant plus le recours aux approches traditionnelles de résumés automatiques, elle offre en revanche l'occasion d'utiliser de nouveaux contextes pour enrichir les résumés, et permet même de créer de nouveaux procédés de résumés automatiques. Par exemple, l'article de (Hu *et al.*, 2007) suggère de procéder au résumé automatique d'une publication tirée d'un blogue en extrayant des phrases représentatives à partir d'informations recueillies de commentaires d'utilisateurs. (Chua et Asur, 2012) se concentrent, quant à eux, sur la corrélation temporelle de gazouillis pour extraire ceux susceptibles d'être pertinents pour le résumé automatique. Enfin, outre le contenu des messages, d'autres approches exploitent les informations liées à l'interaction entre utilisateurs pour produire un résumé des différents échanges (Lin *et al.*, 2009).

– Comme évoqué en section 2.2, l'identification d'événements dans les flux de données est une tâche particulièrement difficile (Allan, 2002). Dans le cadre de l'étude des réseaux sociaux, l'un des défis majeurs est la distinction entre l'information triviale et « polluée » et les événements concrets d'intérêt. La dispersion des

données, l'absence de contexte et la diversité du vocabulaire rendent les techniques traditionnelles d'analyse textuelle difficilement applicables aux gazouillis (Metzler *et al.*, 2007). En outre, différents événements n'atteindront pas la même popularité chez les utilisateurs et peuvent grandement varier sur le plan du contenu, de la période couverte, de la structure inhérente, des relations causales, du nombre de messages générés et du nombre de participants (Nallapati *et al.*, 2004).

– Alors que les médias traditionnels visent, en général, à diffuser une information objective, neutre et factuelle, les médias sociaux sont beaucoup plus porteurs de sentiments voire d'émotion (Neviarouskaya *et al.*, 2011 ; Bringay *et al.*, 2014) (cf. section 2.3). L'information subjective joue donc un rôle essentiel dans l'analyse sémantique des textes issus des réseaux sociaux. L'identification de sentiments repose généralement sur deux familles d'approches. La première est fondée sur des méthodes classiques d'apprentissage supervisé qui proposent des résultats tout à fait satisfaisants pour l'analyse de sentiments (Pang *et al.*, 2002). La seconde s'appuie sur des informations statistiques liées au nombre de descripteurs linguistiques positifs et négatifs qui apparaissent dans chaque texte (Turney, 2002). Dans le cadre de ces approches, il est alors pertinent d'utiliser des ressources existantes telles que SentiWordNet (Esuli et Sebastiani, 2006). Chaque caractéristique de cette ressource est associée à des scores numériques décrivant l'intensité des descripteurs linguistiques selon trois critères : objectif, positif et négatif. Notons que certaines approches récentes se concentrent sur l'identification des émotions associées aux descripteurs spécifiques des réseaux sociaux (par exemple, les hashtags issus des tweets) (Qadir et Riloff, 2014).

4. Conclusion

Les médias sociaux se définissent par le recours à des outils électroniques et à l'Internet dans le but de partager et d'échanger efficacement de l'information et des expériences (Moturu, 2009). Ils donnent accès à une information riche et sans cesse renouvelée que les médias traditionnels ne fournissent pas (Melville *et al.*, 2009). Les deux réseaux sociaux les plus populaires, Facebook et Twitter, sont étudiés dans les articles retenus de ce numéro spécial. Le premier (article de Amitava Das et Björn Gambäck) s'intéresse à l'identification des langues (anglais, bengali et hindi) que nous pouvons retrouver dans une même phrase ou un même message. En effet, comme évoqué dans cet article introductif, l'aspect multilingue associé aux réseaux sociaux demeure une problématique éminemment complexe. Outre le mélange des langues, les messages des réseaux sociaux ont des caractéristiques comme la présence de hashtags qui doivent aussi être étudiés dans les différentes applications. Ces aspects sont pris en compte dans un processus global de détection d'événements proposé dans l'article invité de ce numéro spécial (article de Houssem Eddine Dridi et Guy Lapalme).

Ce numéro spécial montre de quelle manière les méthodes de TAL contribuent à l'analyse des réseaux sociaux. Différents systèmes qui gèrent le contenu des forums

de discussion, des blogues et des microblogues, ont récemment connu des améliorations qui favorisent tant la formation de communautés virtuelles que la connectivité et la collaboration entre les utilisateurs (Osborne *et al.*, 2014). Alors que les médias traditionnels – tels que journaux, télévisions et radios – se caractérisent par un mode de communication unidirectionnel de l'entreprise jusqu'au consommateur, les médias sociaux, eux, proposent différentes plates-formes où l'interaction dans les deux sens est possible. Pour cette raison, ils représentent une source primaire d'information au moment de réaliser une veille stratégique. C'est ainsi que plus récemment, les recherches se sont concentrées sur l'analyse du langage dans les médias sociaux pour comprendre les comportements sociaux et concevoir des systèmes socioadaptés. L'objectif est d'analyser le langage dans une démarche pluridisciplinaire mêlant par exemple, informatique, linguistique, sociolinguistique et psycholinguistique (Brézillon *et al.*, 2013 ; Aiello et McFarland, 2014).

Remerciements

Nous remercions les auteurs pour la qualité des contributions, les relecteurs pour l'évaluation des articles soumis et Jean-Luc Minel pour son soutien et ses conseils avisés tout au long du processus.

5. Bibliographie

Abdaoui A., Azé J., Bringay S., Grabar N., Poncelet P., « Analysis of Forum Posts Written by Patients and Health Professionals », *Proceedings of European Medical Informatics Conference (MIE)*, p. 1185, 2014.

Aiello L. M., McFarland D. A. (eds), *Social Informatics - 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, vol. 8851 of *Lecture Notes in Computer Science*, Springer, 2014.

Allan J. (ed.), *Topic Detection and Tracking : Event-based Information Organization*, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

André P., Bernstein M., Luther K., « Who Gives a Tweet ? : Evaluating Microblog Content Value », *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, p. 471-474, 2012.

Bakliwal A., Foster J., van der Puil J., O'Brien R., Tounsi L., Hughes M., « Sentiment Analysis of Political Tweets : Towards an Accurate Classifier », *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, p. 49-58, June, 2013.

Bouillot F., Hai P. N., Béchet N., Bringay S., Ienco D., Matwin S., Poncelet P., Roche M., Teisseire M., « How to Extract Relevant Knowledge from Tweets ? », *Information Search, Integration and Personalization - International Workshop, ISIP 2012, Springer, Revised Selected Papers*, p. 111-120, 2012.

Brézillon P., Blackburn P., Dapoigny R. (eds), *Modeling and Using Context - 8th International and Interdisciplinary Conference, CONTEXT 2013, Annecy, France, October 28 -31, 2013, Proceedings*, vol. 8175 of *Lecture Notes in Computer Science*, Springer, 2013.

Bringay S., Kergosien E., Pompidor P., Poncelet P., « Identifying the Targets of the Emotions Expressed in Health Forums », *Proceedings of Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, LNCS, Part II*, p. 85-97, 2014.

Chua F. C. T., Asur S., Automatic Summarization of Events from Social Media, Technical report, HP Labs, 2012.

Esuli A., Sebastiani F., « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining », *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, p. 417-422, 2006.

Farzindar A., Gamon M., Inkpen D., Nagarajan M., Danescu-Niculescu-Mizil C. (eds), *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, June, 2013.

Farzindar A., Inkpen D. (eds), *Proceedings of the Workshop on Semantic Analysis in Social Media*, Association for Computational Linguistics, April, 2012.

Farzindar A., Inkpen D., Gamon M., Nagarajan M. (eds), *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Association for Computational Linguistics, April, 2014.

Farzindar A., Wael K., « A Survey of Techniques for Event Detection in Twitter », *Computational Intelligence*, vol. 31, n° 1, p. 132-164, 2015.

Gaio M., Sallaberry C., Nguyen V. T., « Typage de noms toponymiques à des fins d'indexation géographique », *Traitement Automatique des Langues*, vol. 53, n° 2, p. 143-176, 2012.

Hangya V., Berend G., Farkas R., « SZTE-NLP : Sentiment Detection on Twitter Messages », *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, p. 549-553, 2013.

Hu M., Sun A., Lim E.-P., « Comments-oriented Blog Summarization by Sentence Extraction », *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, ACM, p. 901-904, 2007.

Kong L., Schneider N., Swayamdipta S., Bhatia A., Dyer C., Smith N. A., « A Dependency Parser for Tweets », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, p. 1001-1012, 2014.

Lin H., Bilmes J., Xie S., « Graph-based Submodular Selection for Extractive Summarization », *The eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, p. 381-386, 2009.

Melville P., Sindhwani V., Lawrence R. D., « Social Media Analytics : Channeling the Power of the Blogosphere for Marketing Insight », *Proceedings of the Workshop on Information in Networks (WIN)*, 2009.

Metzler D., Dumais S., Meek C., « Similarity Measures for Short Segments of Text », *Advances in Information Retrieval*, vol. 4425 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 16-27, 2007.

Moncla L., Renteria-Agualimpia W., Nogueras-Iso J., Gaio M., « Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus », *Proceedings*

of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 183-192, 2014.

Moturu S., Quantifying the Trustworthiness of User-Generated Social Media Content, PhD thesis, Arizona State University, 2009.

Nallapati R., Feng A., Peng F., Allan J., « Event Threading Within News Topics », *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM)*, ACM, p. 446-453, 2004.

Neviarouskaya A., Prendinger H., Ishizuka M., « Affect Analysis Model : Novel Rule-based Approach to Affect Sensing from Text », *Natural Language Engineering*, vol. 17, n° 1, p. 95-135, January, 2011.

Osborne M., Moran S., McCreadie R., Von Lunen A., Sykora M., Cano E., Ireson N., Macdonald C., Ounis I., He Y., Jackson T., Ciravegna F., O'Brien A., « Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media », *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Association for Computational Linguistics, p. 37-42, 2014.

Pang B., Lee L., Vaithyanathan S., « Thumbs Up ? : Sentiment Classification Using Machine Learning Techniques », *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 79-86, 2002.

Qadir A., Riloff E., « Learning Emotion Indicators from Tweets : Hashtags, Hashtag Patterns, and Phrases », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, p. 1203-1209, 2014.

Sharifi B., Hutton M.-A., Kalita J. K., « Experiments in Microblog Summarization », *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, IEEE, p. 49-56, 2010.

Tang J., Chang Y., Liu H., « Mining Social Media with Social Theories : A Survey », *SIGKDD Explor. Newsl.*, vol. 15, n° 2, p. 20-29, June, 2014.

Turney P. D., « Thumbs Up or Thumbs Down ? : Semantic Orientation Applied to Unsupervised Classification of Reviews », *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Association for Computational Linguistics, p. 417-424, 2002.

Zhou L., Hovy E. H., « On the Summarization of Dynamically Introduced Information : Online Discussions and Blogs. », *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, p. 237, 2006.